

Helsinki 28.3.2000

RECD 28 APR 2000

PO PCT

E T U O I K E U S T O D I S T U S
P R I O R I T Y D O C U M E N T

4



Hakija
Applicant

Alma Media Oyj
Helsinki

Patentihakemus nro
Patent application no

990286

Tekemispäivä
Filing date

12.02.1999

Kansainvälinen luokka
International class

G06F

Keksinnön nimitys
Title of invention

"Mekanismi sähköisen tekstihaun tukemiseksi"

Täten todistetaan, että oheiset asiakirjat ovat tarkkoja jäljennöksiä patentti- ja rekisterihallitukselle alkuaan annetuista selityksestä, patenttivaatimuksista, tiivistelmästä ja piirustuksista.

This is to certify that the annexed documents are true copies of the description, claims, abstract and drawings originally filed with the Finnish Patent Office.

Pirjo Kaila
Tutkimussihteeri

**PRIORITY
DOCUMENT**
SUBMITTED OR TRANSMITTED IN
COMPLIANCE WITH RULE 17.1(a) OR (b)

Maksu 300,- mk
Fee 300,- FIM

Osoite: Arkadiankatu 6 A Puhelin: 09 6939 500 Telefax: 09 6939 5328
P.O.Box 1160 Telephone: + 358 9 6939 500 Telefax: + 358 9 6939 5328
FIN-00101 Helsinki, FINLAND

Mekanismi sähköisen tekstihaun tukemiseksi

Keksinnön tausta

Keksintö liittyy sähköisen tekstihaun tukemiseen, erityisesti kohdistettessa hakuja Internet-tyyppisessä verkossa ja CD-ROM -levyllä julkaistaviin dokumentteihin.

Sähköisessä muodossa julkaistavien dokumenttien määrä ja informaationsältö kasvavat valtavalla nopeudella. Yhä enenevä osa artikkeleista julkaistaan Internetissä tai CD-ROM -levyllä (tai DVD-levyllä).

- Käyttäjä etsii tietoa tällaisista dokumenteista antamalla yhden tai 10 muutaman sanan, joita hän pitää erityisen relevantteina. Näitä sanoja kutsutaan hakusanoiksi. Esimerkiksi maantieliikenneonnettomuksista kiinnostunut käyttäjä voi etsiä hakusanoja "tie", "liikenne", "onnettomuus" jne.

Ohjelmaa, tietokonetta ja palvelua, joka toteuttaa käyttäjän määrittelemän tekstihaun, kutsutaan vastaavasti hakuohjelmaksi, hakukoneeksi ja 15 hakupalveluki. Jos hakuohjelma ensin vastaanottaisi käyttäjän määrittelemät hakusanat ja vasta sitten lähtisi seulomaan koko sen ulottuvilla olevaa informaatiota, haku muodostuisi yleensä toivottoman hitaaksi. Täyneen kirjoitetun CD-ROM -levyn läpikäyminen kestääsi useita minuutteja. Koko Internetin läpikäyminen veisi useita päiviä tai viikkoja. Tämä myös kuormittaisi Internetiä 20 kohtuuttomasti. Koska ensimmäinen haku ei yleensä tuota riittävän hyvää otosta, haku joudutaan toistamaan useita kertoja.

Tämän ongelman ratkaisemiseksi on kehitetty indeksointiohjelmia ja -palveluja, jotka kävät etukäteen läpi käytettävissään olevaa informaatiota ja muodostavat siitä indeksoidun tietokannan, johon voidaan kohdistaa hakuja 25 yhdessä tai muutamassa sekunnissa. Esimerkkejä tällä teknikalla Internetissä toimivista hakupalveluista ovat Lycos, Yahoo ja AltaVista. Esimerkinä omassa tietokoneessa tai lähiverkossa toimivasta hakuohjelmasta olkoon dtSearch, jota valmistaa samanniminen yhtiö. Kaikista näistä on saatavana tietoa www-osoitteella (World Wide Web) www.nimi.com, missä "nimi" on yllä mainittu palvelun tai yhtiön nimi.

Kuvio 1 esittää Internet-tyyppisessä verkossa julkaistavan dokumentin hakua. Internet-tyyppisellä verkolla tarkoitetaan varsinaisen Internet-verkon lisäksi sen suljettuja osa-alueita, joista käytetään nimityksiä intranet, extranet jne. TE (Terminal Equipment) esittää käyttäjän päätelaitteista, jolla 35 tarkoitetaan käyttäjän tietokonetta ja/tai näyttöpäätettä sekä siinä suoritettavaa selainohjelmaa Internet-sivujen esittämiseksi. Viite 1-A esittää hakupalvelun

tarjoajan hakupalvelinta, DNS (Domain Name Server) yhtä nimipalvelun palvelinta eli nimipalvelinta ja viite 1-B Internet-sivuja ylläpitävää WWW-palvelinta eli dokumentin julkaisijan palvelinta.

- Vaiheessa 1-2 käyttäjän antama hakupalvelimen HTTP-muotoinen
- 5 (HyperText Transfer Protocol) Internet-osoite välitetään nimipalvelimelle DNS, joka puolestaan välittää käyttäjälle vaiheessa 1-4 kyseisen hakupalvelimen osoitteen IP-muodossa (Internet Protocol). IP-osoitteenvälisten päätelaitteiston TE muodostaa vaiheessa 1-6 yhteyden hakupalvelimeen 1-A.

- Vaiheessa 1-8 hakupalvelin lähetää WWW-sivunsa käyttäjälle siivunkuvauskielen HTML-muodossa (HyperText Markup Language), ja WWW-sivut esitetään käyttäjän päätelaitteiston TE näytönpäätteellä. Yhteys palvelimen ja päätelaitteiston välillä on auki vain sivun siirtoon kuluvan ajan.
- 10

- Vaiheessa 1-10 käyttäjä antaa hakukoneen hakulomakkeeseen yhden tai useampia hakusanoja, ja ne lähetetään vaiheessa 1-12 hakupalvelimelle. Hakuohjelma etsii vaiheessa 1-14 kyseisiä hakusanoja hakupalvelimen tietokannasta. Lista löytyneistä, hakusanat sisältävistä dokumenteista palautetaan käyttäjälle vaiheessa 1-16.
- 15

- Vaiheessa 1-18 käyttäjä voi selata löytyneitä dokumentteja Internet-selaimessaan. Kun hän haluaa tutustua johonkin haussa löytyneeseen dokumenttiin, hän antaa kyseisen dokumentin WWW-osoitteen selaimelleen (esimerkiksi valitsemalla sen hakupalvelun tuottamasta listasta), joka ottaa yhteyden vaiheessa 1-20 nimipalvelimeen. Tämä palauttaa vaiheessa 1-22 kyseisen dokumentin IP-osoitteen selaimelle, joka tämän IP-osoitteen avulla pyytää kyseistä dokumenttia vaiheessa 1-24. Vaiheessa 1-26 kyseinen WWW-sivu lähetetään käyttäjälle. Mikäli käyttäjä haluaa selata muita hakukoneen löytämiä dokumentteja, hän voi vaiheessa 1-28 palata takaisin hakukoneen listaukseen löytyneistä dokumenteista.
- 20
 - 25

- Käyttäjä voi toistaa vaiheita 1-18 ... 1-28, kunnes hän on käynyt läpi kaikki hakukoneen löytämät dokumentit ja/tai kunnes hän haluaa lopettaa dokumenttien selaamisen.
- 30

- Jotta edellä kuvattu haku olisi tehokasta, eri hakupalveluilla on erilaisia tekniikoita hakusanojen yhdistelemiseksi. Yleisesti käytetään loogisia operaattoreita AND, OR ja NOT sekä sulkumerkkejä. Esimerkiksi haku sanoilla "tie AND onnettomuuus" etsii dokumentteja, joissa esiintyvät sanat "tie" ja "onnettomuuus".
- 35

Haku ei kuitenkaan yleensä tuota relevanttia tietoa, mikäli annetaan vain hakusanojen yhdistelmä. Sen vuoksi useimmat hakupalvelut tunnistavat myös läheisyysoperaattorin. Altavistan tapauksessa tämä on NEAR. Haku sanoilla "tie NEAR onnettamuus" etsii dokumentteja, joissa sanat "tie" ja "onnettamuus" esiintyvät korkeintaan 10 sanan etäisyydellä toisistaan. Myös 5 dtSearch -ohjelmalla voidaan määritellä maksimaalinen sanojen etäisyys: läheisyysoperaattori w/n, missä n=1, 2, ... , edellyttää että sanat esiintyvät korkeintaan n:n sanan etäisyydellä toisistaan.

Tämän keksinnön perustana oleva ongelma on, että etukäteen tapahtuva indeksointi toimii huonosti kielissä, joissa sanoilla on useita taivutusmuotoja. Suomen kielen nomineilla ja verbien nominaalimuodoilla on 15 sijamuotoa, unkari kielessä peräti 21. Kun otetaan huomioon yksikkö- ja monikkomuodot, possessiivisuffiksit ja muut päätteet, mahdollisia taivutusmuotoja on useita satoja.

15 Taivutettujen muotojen löytämiseksi kaikki yllä mainitut hakupalvelut tukevat villi- eli tähtimerkin (*) käyttöä: tähtimerkkiä voidaan käyttää osoittamaan, että sanan loppu on katkaistu ja hakupalvelun tulee löytää kaikki sanat, jotka alkavat annetulla tavalla. Esimerkiksi annettaessa hakusanaksi "onnettamuu" tulee hakupalvelun löytää sanat "onnettamuudet", 20 "onnettamuuksista", "onnettamuustutkintalautakunta" jne.

Tähtimerkin käytössä on kuitenkin ongelmia ja rajoituksia. Esimerkiksi AltaVista -hakupalvelu vaatii, että hakusanasta annetaan ainakin kolme kirjainta ennen tähtimerkkiä. Kuitenkin esimerkiksi sanan "tie" taipumaton osa on vain yhden kirjaimen pituinen: "teiden", "teillä" jne. Toinen ongelma on, että hakusanalla "tie*" palautetaan kaikki tie-alkuiset sanat, kuten "tiede", "tietokone", "tietoliikenne", "tietysti", "tienoo" ja "tietoisuus", kaikissa taivutusmuodoissaan. Lyhytvartaloisten sanojen etsiminen tunnetulla teknikalla tuottaa siis erittäin epärelevanttia tai ei lainkaan tietoa.

Keksinnön lyhyt selostus

30 Keksinnön eräänä tarkoituksena on kehittää sellainen sähköisen dokumentin rakenne, jolla dokumentin indeksinnin jälkeen ei esiinny yllä mainittuja ongelmia. Toisella tavalla nähtynä eksinnön tavoitteena on kehittää menetelmä ja laitteisto tällaisten dokumenttien tuottamiseksi. Keksinnön tavoitteet saavutetaan menetelmällä ja järjestelmällä, joille on tunnusomaista se, 35 mitä sanotaan itsenäisissä patenttivaatimuksissa. Keksinnön edulliset suoritusmuodot ovat epäitsenäisten patenttivaatimusten kohteena.

Keksintö perustuu siihen, että sähköisesti julkaistavaa dokumenttia täydennetään lisäämällä siihen dokumentin tekstiosuuden sisältämät sanat perusmuodoissaan ja alkuperäisessä järjestyksessä. Sanojen lisääminen perusmuodoissaan saa aikaan sen, että hakupalvelu löytää seuraavan indeksoinnin 5 jälkeen keksinnön mukaisesti täydennetyn dokumentin, vaikka alkuperäisessä dokumentissa sana ei esiintyisi lainkaan perusmuotoisena.

Jäljempänä käytetään nimitystä "täydennysosa" siitä osasta, joka sisältää keksinnön mukaisesti lisättyt sanat. Vastaavasti "perusosa" on se osa, joka sisältää alkuperäisen dokumentin.

10 Itse asiassa on tunnettua lisätä dokumentteihin käsin joitakin perusmuotoisia avainsanoja. Tieteellisten dokumenttien otsikon alla tai vaihtoehtoisesti dokumentin lopussa käytetään joskus kenttää "avainsanat", jossa esiintyy muutama avainsana. Tämä ei kuitenkaan ratkaise ongelmaa toivotulla tavalla, koska perusmuotoisten avainsanojen määrä on hyvin rajallinen, eikä 15 läheisyysoperaattori toimi oikein. Pitkässä artikkelissa voidaan puhua useasta täysin erillisestä asiasta, mutta avainsanakentässä vastaavat hakusanat ovat kuitenkin lähellä toisiaan.

20 Keksinnön mukainen teknikka, jossa sanat lisätään alkuperäisessä järjestyksessä saa aikaan sen, että hakupalvelu osaa käyttää oikein läheisyysoperaattoreita. Esimerkiksi haku sanoilla "tie NEAR onnettomuus" löytäisi dokumentin, joissa esiintyy tekstifragmentti "teillämme tapahtuneet onnettomuudet", vaikka dokumentti ei sisältäisi lainkaan sanoja "tie" tai "onnettomuus" perusmuodoissaan.

25 Koska dokumentin tekstiosan sanat lisätään alkuperäisessä järjestyksessä, näyttäisi siltä että dokumentin pituus likimain kaksinkertaistuu. Tämä pitää paikkansa vain tekstiä sisältävien dokumenttien suhteen. Useimpiin dokumentteihin liittyy kuitenkin kuvia, joiden vaatima muistitila ylittää moninkertaisesti tekstiosuuden vaatiman muistilan, joten tekstiosuuden kaksinkertais-taminen ei merkittävästi kasvata koko dokumentin vaatimaa muistitilaa.

30 Muistilan vähäisen kasvamisen vastapainoksi keksinnön mukainen teknikka tuo vielä yhden yllättävän edun: näin täydennettyjen dokumenttien relevanssi kasvaa näennäisesti ainakin kaksinkertaiseksi, koska dokumenteissa on käyttäjän valitsemia hakusanoja kaksinkertainen määrä. Keksinnön mukaisesti täydennetyn dokumentin julkaisija saa siis sanomansa paremmin pe-rille. Dokumentin relevanssi kasvaa kaksinkertaiseksi sellaisten hakusanojen 35 suhteen, joilla on niin pitkä vartalo, että niitä voidaan iuotettavasti hakea tähti-suhteeksi.

merkillä, esimerkiksi "onnettomuu". Lyhytvartaloisten sanojen kohdalla, joita tunnetulla tekniikalla ei voida hakea lainkaan, dokumentin relevanssi kasvaa moninkertaiseksi, mikä johtuu siitä, että tunnetulla tekniikalla tällaiset dokumentit eivät hakupalvelulle ole lainkaan relevantteja. (Ne voivat olla osittain

- 5 relevantteja siinä tapauksessa, että käyttäjä antaa useita hakusanoja, joista muut sanat ovat sellaisia, että hakupalvelu löytää ne.) Tässä kappaleessa relevanssilla ei siis tarkoiteta sitä, kuinka relevantti jokin dokumentti on käyttäjälle, mikäli hän sen löytäisi, vaan sillä tarkoitetaan hakupalvelun tuottamaa mittalukua, jonka laskenta perustuu siihen, kuinka monta annetuista hakusanoista esiintyy dokumentissa, ja mahdollisesti kuinka usein ne esiintyvät.

- 10 Dokumentin käyttäjät (henkilöt, jotka etsivät kyseistä dokumenttia) eivät voi etukäteen tiedää, mitkä dokumentit on täydennetty keksinnön mukaisella tavalla, ja mitkä eivät ole. Myös tästä syystä sanojen lisääminen alkuperäisessä järjestysessä on erittäin tärkeä ominaisuus, koska käyttäjien ei tarvitse muuttaa hakutottumuksiaan mitenkään, vaan he voivat käyttää läheisyyssoperaattoria totutulla tavalla.

Käyttäjän hakutoiminto ei kuitenkaan lopu siihen, että hakupalvelu löytää hänelle jonkin hakusanat sisältävän dokumentin. Hänen on yleensä vielä löydettävä relevantit alueet dokumentin sisältä.

- 15 20 Oletetaan aluksi, että dokumentit täydennetään yksinkertaisesti lisäämällä perusmuotoiset sanat dokumentin loppuun. Käyttäjä voi etsiä tästä täydennysosuudesta perusmuotoisia hakusanoja selain- tai tekstinkäsittelyohjelman hakutoiminoilla. Mikäli hakusana on lyhytvartaloinen, käyttäjä ei voi etsiä sitä dokumentin perusosasta, mutta hän voi katsoa täydennysosasta jonkin hakusanan lähellä olevan pidemmän ja harvinainen sanan, ja etsiä sen dokumentin perusosasta. Tässä suhteessa keksinnön mukainen tekniikka voi aiheuttaa pieniä muutoksia käyttäjän toimintatapoihin, mutta muutos näkyy vasta sitten kun hakuohjelma on jo löytänyt dokumentin ja käyttäjä selaa sitä. Mikäli dokumentti on lyhyt, tai käyttäjä muusta syystä päättää lukea sen kokonaan, käyttäjä ei joudu muuttamaan toimintatapojaan.

- 25 30 35 Dokumentin perusmuotoisten sanojen lisääminen dokumentin loppuun vääristää dokumentin ulkoasua. Teksti näyttää sellaisen ihmisen kirjoitamalta, joka ei ymmärrä kielestä mitään, vaan kääntää koneellisesti sanakirjan avulla. Dokumentin kirjoittajan mielestä tällaista voitaisiin pitää jopa respektioikeuden loukkauksena. (Respektioikeus tarkoittaa, että kaupallisesta leviysoikeudesta riippumatta teosta ei saa esittää loukkaavalla tavalla.) Sen

vuoksi täydennysosa on edullista liittää dokumenttiin tavalla, joka estää sen näkymisen dokumentin normaalikäytössä. Esimerkiksi HTML-koodattuun (HyperText Markup Language) dokumenttiin voidaan liittää ainakin yksi kommentti- tai metakoodikenttä, joka sisältää keksinnön mukaisen täydennysosan.

- 5 Vaihtoehtoinen tapa on yhden tai useamman kuvan lataaminen täydennysosan päälle. Kun käyttäjä haluaa etsiä hakusanaa tästä täydennysosasta, hän avaa dokumentin selainohjelmallaan ja näyttää dokumentin sisältämät HTML-kieliset käskyt. Esimerkiksi Internet Explorer -ohjelmalla tämä tapahtuu käsikyllä View/Source. Vastaavasti kehittyneillä tekstinkäsittelyohjelmilla on mahdollista asettaa täydennysosalle attribuutti "piiloteksti", jolloin se saadaan näkyviin näyttämällä normaalisti näkymättömät ohjaus- ja erikoismerkit.

Tunnetut hakuteknikat eivät löydä hakusanoja, jotka esiintyvät yhdyssanan osina, mutta eivät sen alussa. Läheisyysoperaattori ei myöskään toimi, mikäli hakusanat esiintyvät yhdyssanan eri osina. Esimerkiksi sanasta "maantieliikenteen" ei löydetä sanoja "tie" eikä "liikenne" eikä varsinkaan näitä sanoja lähellä toisiaan. Sen vuoksi erään toisen edullisen suoritusmuodon mukaan keksinnön mukainen täydennysosa sisältää kunkin yhdyssanan kohdalla kyseisen yhdyssanan perusmuodon lisäksi yhdyssanan osien perusmuodot erillisinä sanoina. Esimerkiksi taivutetussa muodossa olevan yhdyssanan "maantieliikenneonnettомуksien" kohdalla täydennysosa sisältäisi sanat "maantieliikenneonnettомуus", "maa", "tie", "liikenne" ja "onnettомуus". Näin täydennetty dokumentti löytyy, mikäli käyttäjä hakee sanoja "tie" ja "liikenne", jopa silloin kun käyttäjä vaatii, että nämä sanat esiintyvät lähekkäin.

Vielä erään edullisen suoritusmuodon mukaan keksinnön mukainen täydennysosa sisältää perusmuotoisen yhdyssanan ja sen osien lisäksi kaikki yhdyssanan osien yhdistelmät siten, että yhdyssanan muut kuin viimeinen osa ovat siinä muodossa kuin ne esiintyvät dokumentissa ja yhdyssanan viimeinen osa on perusmuodossaan. Yhdyssanan osien yhdistelmät ovat lisäksi alkuperäisessä järjestyksessään, siis edellisen esimerkin tapauksessa sanat "maantie", "tieliikenne" ja "liikenneonnettомуus" sekä "maantieliikenne", ja "tieliikenneonnettомуus".

Vielä erään edullisen suoritusmuodon mukaan keksinnön mukainen täydennysosa lisätään keskitetyssä palvelimessa, jotta jokaisen dokumentin julkaisijan ei tarvitse hankkia ohjelmistoa, joka osaa muuntaa sanoja perusmuotoonsa. Alkuperäinen dokumentti voidaan lähetä täydennystä varten levylle, sähköpostin liitetiedostona, Internetin FTP-protokoilialla tms.

Kuvioiden lyhyt selostus

Keksintöä selostetaan nyt lähemmin edullisten suoritusmuotojen yhteydessä, viitaten oheisiin piirroksiin, joista:

- 5 Kuvio 1 on yhdistetty vuo- ja signalointikaavio, joka esittää Internet-tyypisessä verkossa julkaistavan dokumentin hakua;
- 6 Kuvio 2 esittää signalointikaaviota keksinnön mukaisen täydennysosan lisäämiseksi;
- 7 Kuvio 3A esittää esikäsitellyn dokumentin ja täydennetyn dokumentin rakenteita.
- 10 Kuvio 3B esittää täydennetyn dokumentin rakennetta, missä täydennysosan päälle on ladattu kuva.

Keksinnön yksityiskohtainen selostus

- 15 Eräs mahdollinen tekniikka keksinnön mukaisen täydennysosan lisäämiseksi dokumenttiin esitetään kuviossa 2, jossa viite 2-A esittää dokumentin julkaisijan palvelinta, DNS nimipalvelinta, viite 2-B edellä mainittua keskitettyä palvelinta eli täydennyspalvelun tuottajan palvelinta ja viite 2-C hakupalvelun tarjoajan palvelinta

- 20 Dokumentin julkaisijalla tarkoitetaan sitä, joka haluaa julkaista keksinnön mukaisesti täydennetyn dokumentin. Täydennyspalvelun tuottaja puolestaan tarjoaa keksinnön mukaisen palvelun täydennysosan lisäämiseksi dokumenttiin.

- 25 Vaiheessa 2-10 dokumentin julkaisija lähettilä täydennyspalvelun tuottajan palvelimen WWW-osoitteeseen nimipalvelimelle DNS, joka palauttaa vaiheessa 2-12 vastaan IP-osoitteeseen dokumentin julkaisijalle. Tämän avulla dokumentin julkaisija pääsee vaiheessa 2-14 täydennyspalvelun tuottajan Internet-sivulle. Vaiheessa 2-16 dokumentin julkaisijan selainohjelma noutaa täydennyspalvelun tuottajan WWW-sivun-/sivut päätelaitteelleen.

- 30 Kyseisellä WWW-sivulla voidaan esittää ainakin sähköpostiosoite, jonne dokumentin julkaisija voi lähettilä dokumentin keksinnön mukaista täydennystä varten. Sivuilla täydennyspalvelun tuottaja voi kertoa esimerkiksi tarjoamastaan palvelusta ja antaa ohjeita sen käyttämiseksi.

- 35 Saatuaan sähköpostiosoitteen tietoansa dokumentin julkaisija voi vaiheessa 2-18 lähettilä dokumentinsa muokattavaksi täydennyspalvelun tuottajalle liittämällä sen esimerkiksi sähköpostin liitetiedostoksi (attachment).

Toinen mahdollinen dokumentin lähetystapa on FTP-siirto (File Transfer Protocol). Dokumentin siirtotapa ei kuitenkaan ole keksinnön kannalta oleellinen.

Vaiheessa 2-20 dokumentin julkaisijan HTML-muotoista dokumenttia muokataan täydennyspalvelun tuottajan palvelimella: siihen lisätään keksinnön mukainen täydennysosa. Tämän jälkeen kyseinen täydennetty dokumentti palautetaan dokumentin julkaisijalle vaiheessa 2-22 joko sähköpostilla tai FTP:n avulla. Sen jälkeen kun hakupalvelu on indeksoinut vaiheessa 2-24 kyseisen dokumentin, dokumentin käyttäjä voi etsiä kyseistä dokumenttia myös perusmuotoisilla sanoilla ja, mikäli yhdyssanat jaetaan osiin, myös niiden 10 osien perusmuodoilla. Internet-tyyppisessä verkossa julkaistavan dokumentin hakua kuvataan kuvion 1 vaiheesta 1-12 alkaen.

Hakupalvelin on konfiguroitavissa siten, että täydennysosan perusmuotoisille sanoille voidaan antaa myös enemmän painoarvoa eli relevanssipisteitä kuin normaalisti. Tämä tarkoittaa sitä, että dokumentit, joiden sisältämällä sanoilla on enemmän relevanssipisteitä sijoitetaan hakutulokset esittävässä listassa lähemmäksi listan kärkipäätä kuin dokumentit, joiden sisältämällä sanoilla on vähemmän relevanssipisteitä. Jos täydennysosan perusmuotoisille sanoille ei anneta ollenkaan relevanssipisteitä tai jos sanat jätetään indeksoimatta, dokumenttia ei löydetä täydennysosan avulla.

Kuvio 3A esittää esikäsityllyn dokumentin 3-2 ja täydennetyn dokumentin 3-20 rakenteita. Alkuperäinen dokumentti voi olla esimerkiksi tekstinäkäsitteilyohjelmalla kirjoitettu tekstisivu muotoiluineen. Esikäsitylty dokumentti on alkuperäinen dokumentti muokattuna esimerkiksi HTML-kieliseksi. Dokumenteissa voi olla lisäksi kuvia, taulukoita, kehyksiä ja/tai muita Internetsivuilla tallennettavissa olevia objekteja. Viite 3-10 kuvaa HTML-kielen aloitusmerkkiä <HTML> ja viite 3-12 kuvaa HTML-kielen lopetusmerkkiä </HTML>. Kyseisten merkkien välissä on dokumentin sisältö 3-4.

Keksinnön mukaisesti täydennetty dokumentti 3-20 sisältää myös aloitusmerkin 3-10 ja lopetusmerkin 3-12 sekä dokumentin sisällön 3-4. Tämän lisäksi täydennettyn dokumenttiin 3-20 on lisätty täydennysosa 3-24, jossa kaikki dokumentissa esiintyvät sanat ovat perusmuodoissaan alkuperäisessä järjestyksessä. Täydennysosan koodaus voi tapahtua esimerkiksi koodaamalla se metakoodiksi (Metakeyword) tai HTML-kommentiksi. HTML-tiedosto voi sisältää useita HTML-komentteja. HTML-tiedoston eritinmerkkinä toimivat "<! ..." ja "...>". Kommentin sijainnilla tiedostossa ei ole merkitystä. Kommentti voi olla dokumentin (3-20) alussa, lopussa tai sen keskellä. Näiden

- tekniikoiden sijasta tai niiden lisäksi täydennysosan 3-24 päälle voidaan ladata yksi tai useampi kuva. Tätä täydennetyn dokumentin 3-40 rakennetta esitetään kuviossa 3B. Kun täydennysosa 3-24 on kuvan 3-44 alla, täydennysosa ei tule näkyviin dokumentin normaalikäytössä. Tällöin näkyvillä on ainoastaan esikäsitetty dokumentti 3-42.
- Perusmuotoisten sanojen lisäksi dokumentin täydennysosaan voidaan lisätä myös sanojen eri variaatioita, synonyymejä ja rinnakkaismerkityksiä. Tällöin dokumentin relevanssi kasvaa edelleen, koska dokumenttia voidaan etsiä myös hakusanoilla, joita ei esiinny alkuperäisessä dokumentissa.
- Alan ammattilaiselle on ilmeistä, että tekniikan kehittyessä keksinnön perusajatus voidaan toteuttaa monin eri tavoin. Keksintö ja sen suoritusmuodot eivät siten rajoitu yllä kuvattuihin esimerkkeihin vaan ne voivat vaihdella patenttivaatimusten puitteissa.

Patenttivaatimukset

1. Menetelmä ainakin tekstiosuuden sisältävän dokumentin (3-20, 3-40) julkaisemiseksi yhdelle tai useammalle käyttäjälle, jossa menetelmässä:

- dokumentin (3-20, 3-40) tosiaikaisen haun tehostamiseksi dokumenttiin (3-20, 3-40) kohdistetaan ainakin yksi indeksointi (2-24) ja

5 - indeksoinnin tulos tallennetaan,

tunnettu siitä, että ennen mainittua ainakin yhtä indeksointia (2-24) dokumenttia (3-20, 3-40) täydennetään lisäämällä (2-20) siihen täydenyysosa (3-24), joka sisältää olennaisesti ainakin kyseisen dokumentin (3-20, 3-40) tekstiosuuden sisältämät sanat perusmuodoissaan alkuperäisessä järjestyksessä.

10 2. Patenttivaatimuksen 1 mukainen menetelmä, tunnettua siitä, että täydenyysosa (3-24) liitetään dokumenttiin (3-20, 3-40) tavalla, joka estää sen näkymisen dokumentin normaalikäytössä.

15 3. Patenttivaatimuksen 1 mukainen menetelmä, tunnettua siitä, että täydenyysosa (3-24) sisältää kunkin yhdyssanan kohdalla kyseisen yhdyssanan perusmuodon lisäksi yhdyssanan osien perusmuodot erillisinä sanoina.

20 4. Patenttivaatimuksen 3 mukainen menetelmä, tunnettua siitä, että täydenyysosa (3-24) sisältää lisäksi kaikki yhdyssanan osien yhdistelmät, missä osat ovat alkuperäisessä järjestyksessä.

25 5. Patenttivaatimuksen 1 mukainen menetelmä, tunnettua siitä, että useita dokumentteja julkaistaan usealla julkaisupalvelimella ja että täydenyysosa (3-24) lisätään täydenyspalvelimella (2-B), joka on yhteinen usealle julkaisupalvelimelle.

6. Patenttivaatimuksen 5 mukainen menetelmä, tunnettua siitä, että täydenyspalvelin (2-B) vastaanottaa (2-18) ja lähettää (2-22) täydennetävät dokumentit (3-20, 3-40) IP-protokollaa käyttävän tietoliikenneverkon kautta.

30 7. Laitteisto (2-B) elektronisen tekstihaun tukemiseksi, joka laitteisto (2-B) on sovitettu vastaanottamaan ainakin tekstiosuuden sisältämän dokumentin (3-20, 3-40), tunnettua siitä, että dokumentin (3-20, 3-40) tosiaikaisen haun tehostamiseksi laitteisto (2-B) on sovitettu lisäämään täydenyysosan (3-24), joka sisältää olennaisesti ainakin kyseisen dokumentin (3-20, 3-40) tekstiosuuden sisältämät sanat perusmuodoissaan alkuperäisessä järjestyksessä.

8. Patenttivaatimuksen 7 mukainen laitteisto (2-B), t u n n e t t u siitä, että laitteisto (2-B) on sovitettu vastaanottamaan (2-18) ja lähetämään (2-22) dokumentti (3-20, 3-40) IP-protokollaa käyttävän tietoliikenneverkon kautta.

- 5 9. Järjestely dokumenttien julkaisemiseksi IP-protokollaa käyttävän tietoliikenneverkon kautta, joka järjestely käsittää ainakin yhden julkaisupalvelimen (1-B, 2-A) mainitun dokumentin julkaisemiseksi, ainakin yhden indeksointipalvelimen (1-A, 2-C) mainitun dokumentin indeksoimiseksi ja ainakin yhden päätelaitteen (TE) kyselyn lähetämiseksi mainitulle ainakin yhdelle indeksointipalvelimelle (1-A, 2-C), t u n n e t t u siitä, että dokumentin tosiaikaisen haun tehostamiseksi järjestely lisäksi käsittää patenttivaatimuksen 7 tai 8 mukaisen laitteiston (2-B).
- 10 10. Sähköisessä muodossa julkaistava dokumentti (3-20, 3-40), joka sisältää ainakin tekstiosuuden, t u n n e t t u siitä, että dokumentin (3-20, 3-40) tosiaikaisen haun tehostamiseksi dokumentti (3-20, 3-40) käsittää täydennysosan (3-24), joka sisältää olennaisesti ainakin kyseisen dokumentin (3-20, 3-40) tekstiosuuden sisältämät sanat perusmuodoissaan alkuperäisessä järjestyksessä.

(57) Tiivistelmä

Sähköisesti julkaistavaan dokumenttiin (3-2) lisätään täydennysosana (3-24) ainakin dokumentin (3-2) tekstiosuuden sisältämät sanat perusmuodoissaan, alkuperäisessä järjestyksessä ja kunkin yhdyssanan kohdalla lisäksi yhdyssanan osien perusmuodot erillisinä sanoina. Täydenyssanan lisääminen (2-20) saa aikaan sen, että hakupalvelu (1-A, 2-C) löytää seuraavan indeksoinnin (2-24) jälkeen keksinnön mukaisesti täydennetyn dokumentin (3-20, 3-40), vaikka alkuperäisessä dokumentissa sana ei esiintyisi lainkaan perusmuotoisena.

(Kuvio 3A)

FIG. 7

1/3

1-4

HAKUVAJUEEN

TE

DNS

DOKUMENTIN
JULKAISJA

1-6

1-8

IP - OSOITE
WWW-SIVUT

1-2

WWW-OSOITE

1-4

IP - OSOITE

ANNA HAKUSANA

1-4

1-72

HAKUSANA(T)

-ETS- DOKUMENTIT

1-16 LÖYDETYT DOKUMENTIT

7-38

7-78

SEKÄÄ DOKUMENTTIA

1-20

WWW-OSOITE

1-22

IP - OSOITE

1-24

IP - OSOITE

1-26

WWW-SIVUT

LOPPU

FIG. 2

2-C

2-A

2-B

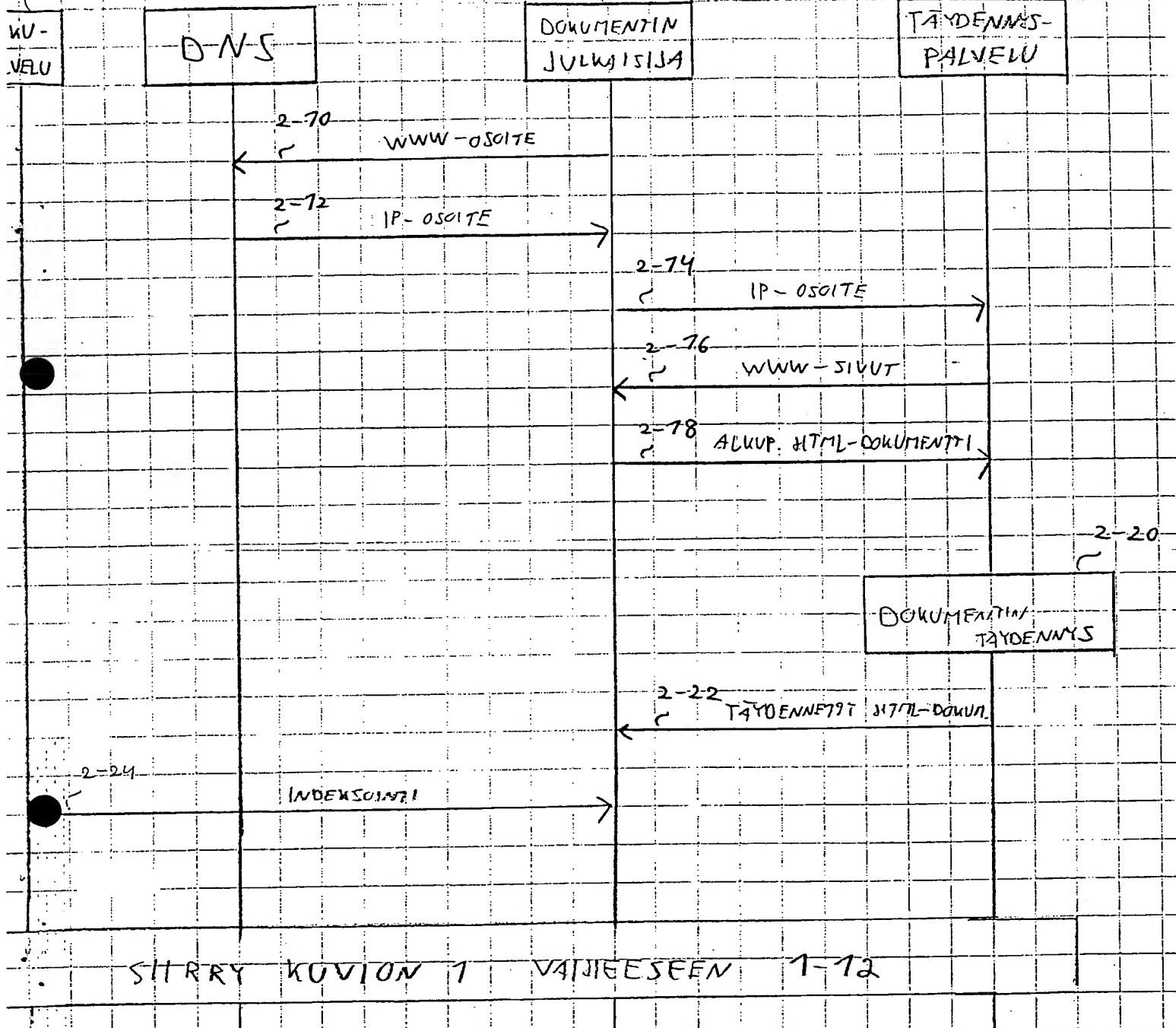


FIG. 3A

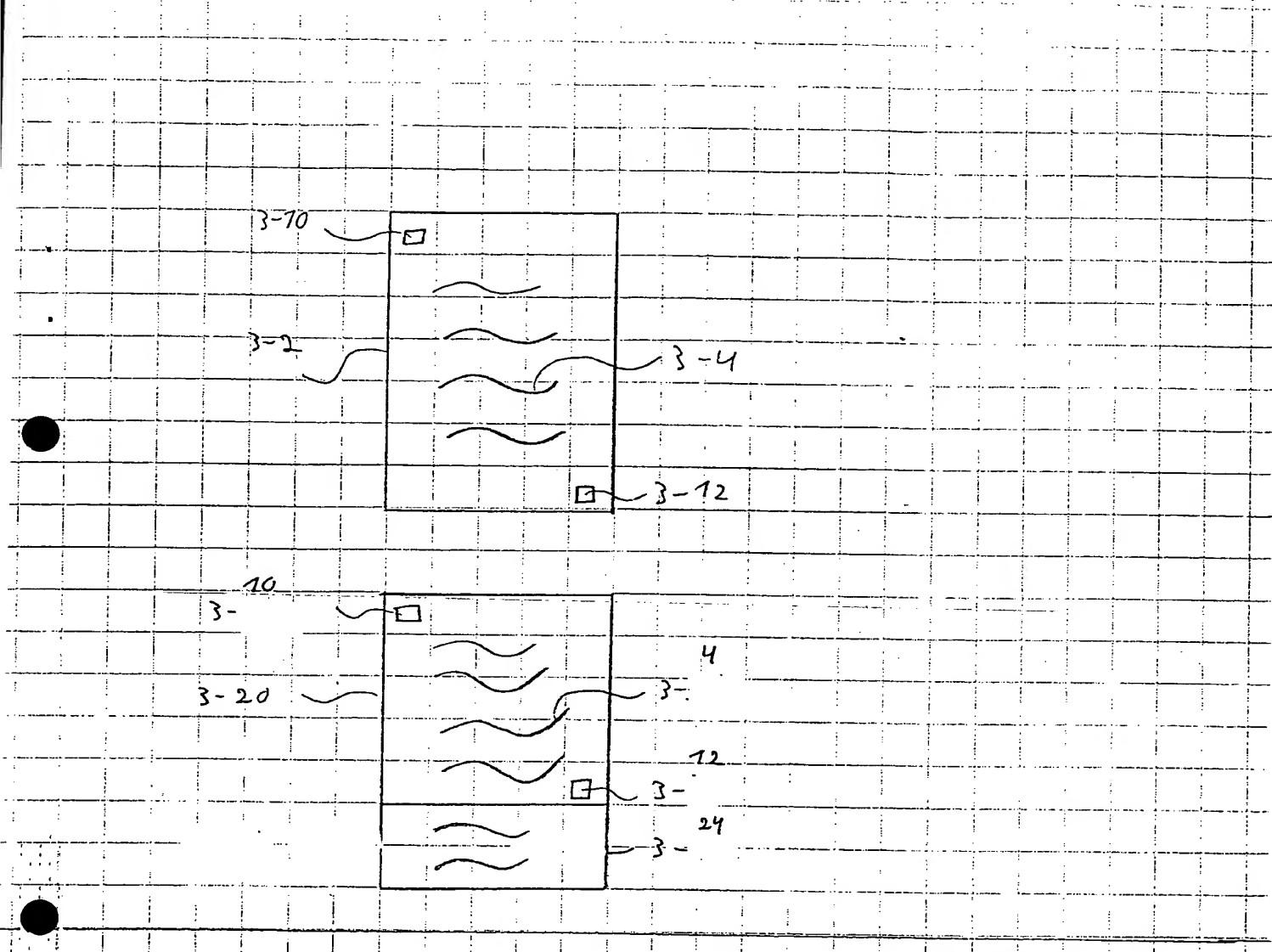


FIG. 3B

